Design Methods for Artificial Intelligence Fairness and Transparency

Simone Stumpf[a], Lorenzo Strappelli[a], Subeida Ahmed[a], Yuri Nakao[b], Aisha Naseer[c], Giulia Del Gamba[d] and Daniele Regoli[d]

[a]City, University of London, Northampton Square, London, UK
[b]Fujitsu Laboratories Ltd., Kawasaki, Japan
[c]Fujitsu Laboratories of Europe, Hayes, UK
[d]Intesa Sanpaolo S.p.A., Turin, Italy

Abstract

Fairness and transparency in artificial intelligence (AI) continue to become more prevalent as topics for research, design and development. General principles and guidelines for designing ethical and responsible AI systems have been proposed, yet there is a lack of design methods for these kinds of systems. In this paper, we present CoFAIR, a novel method to design user interfaces for exploring fairness, consisting of series of co-design workshops, and wider evaluation. This method can be readily applied in practice by researchers, designers and developers to create responsible and ethical AI systems.

Keywords

fairness, transparency, explanations, design, methods

1. Introduction

There has been extraordinary interest in making artificial intelligence (AI) systems ethical and responsible over the last decade [1, 2]. Many principles and guidelines have been proposed to ensure considerations for fairness, accountability, and transparency are made in the design and development of these systems [3]. However, these guidelines are fairly abstract and do not lend themselves to guiding how and what to design. Recent work [4] has started to investigate design patterns to guide detailed user interface design. Overarching design methods for designing transparent AI systems, beyond the User-Centred Design (UCD) process, have also been proposed [5].

In this paper, we review existing work on design methods to guide designers of responsible and ethical AI systems and user interfaces. We then present a new method, Co-designing Fair AI InteRactions (CoFAIR), which consists of a series of co-design workshops followed by a broader evaluation, to create suitable user interfaces that lend themselves to exploring fairness by targeted user groups. We show the application of this method through a case study. We discuss the limitations of our approach, and how this method might be generalised to designing for ethical and responsible AI systems.

Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA
Simone.Stumpf.1@city.ac.uk (S. Stumpf); Lorenzo.Strappelli@city.ac.uk (L. Strappelli); Subeida.Ahmed@city.ac.uk (S. Ahmed); nakao.yuri@fujitsu.com (Y. Nakao); Aisha.Naseer@uk.fujitsu.com (A. Naseer); giulia.delgamba@intesasanpaolo.com (G.D. Gamba); daniele.regoli@intesasanpaolo.com (D. Regoli)
0000-0001-6482-1973 (S. Stumpf); 0000-0002-6813-9952 (Y. Nakao); 0000-0003-2711-8343 (D. Regoli)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

It has been realised that Artificial intelligence and machine learning pose unique design challenges that merit new design practices [6, 7, 8, 9]. In the last few years, a number of approaches have been suggested to ease the design and development of responsible and ethical AI systems. Here, we present an overview of guidelines to designing ethical AI systems, before turning to describing work that aims to address design patterns and methods.

2.1. Design Guidelines

Considerable thought has been given to providing guidelines for designing and developing these ethical AI systems. The most well-known of these have been developed by Microsoft, Google and IBM, with some efforts also being produced by the High-Level Expert Group (HLEG) on AI set up by the European Commission. We will briefly review these efforts but see [3] for a comprehensive survey of AI ethics guidelines.

Microsoft's Guidelines for Human-AI Interactions [10] as part of their Responsible AI area are implemented as a set of eighteen cards. Each card describes a guideline and some examples of how that guideline might apply in practice, over four stages of use: 'initially', 'during interaction', 'when wrong', and 'over time'. These guidelines provide designers and developers with high-level considerations to make during the design process. For example, guideline 6 prompts to "mitigate social biases" during interaction by ensuring that "the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases." Guideline 11 is to "make clear why the system did what it did" when wrong and suggests to "enable the user to access an explanation of why the AI system behaved as it did". While each comes with an example of how this might be realised in practice, it is up to the designer or developer to craft appropriate ways to implement this guideline.

Google's Responsible AI practices [11, 12] suggest that ethical AI systems should be designed following best practices for software systems but then supplemented with considerations specific to machine learning. Overall a human-centered design approach shoudl be followed to actively consider fairness, interpretability, privacy and security from the outset. Specific advice for designing the user experience for AI systems has been given by the People + AI Handbook [12], such as identifying user needs and their mental models, or addressing explainability and trust. While these guidelines do not explicitly surface fairness as a specific consideration, it is covered when collecting and evaluating data and also in communicating with users.

IBM's Everyday Ethics for Artificial Intelligence [13, 14] suggests five areas to focus on in the development of ethical AI systems: accountability, value alignment, explainability, fairness and user data rights. The guidelines present a rationale of why these aspects require attention, make recommendations for actions to take and for questions the design team should consider, and provide examples of implementations.

The HLEG on AI ethics guidelines for trustworthy AI [15] set out a framework for ethical principles and associated requirements that should be covered in AI development. In applying this framework, the report suggests adopting both technical and non-technical methods, such as transparency-by-design or inclusive design teams. In order to assess that AI has been developed in accordance with these principles and requirements, the report also puts forward a checklist to be used within design practices.

While guidelines to develop responsible and ethical AI have some use to stimulate discussions within design teams about high-level

concepts and requirements that need to be met, as noted previously [16], these guidelines are fairly abstract and are difficult for designers and developers to implement into practice.

2.2. Design Patterns

Currently, there is a lack of design patterns for AI systems, which tells designers and developers *what* to design. In HCI and data visualisation, design patterns for common use cases and scenarios on well-studied technologies are readily available[1]. These tell designers and developers how to support interactions and communications through a user interface. Similarly, there has been a line of research in Explainable AI (XAI) that aims to establish what information to communicate and what interactions to support in order to make a system transparent. High-level principles for explainability and controllabity have been proposed [17], such as 'be sound', 'be complete', 'be actionable', and 'be reversible'.

In addition, there is a emerging body of research that aims to investigate what is most effective in terms of user interfaces that provide explanations. A lot of work has focused on what information should be available to users and how this information should be communicated via text, graphics or visualizations [18, 19, 20, 21, 14]. A recent effort to start developing design patterns [4], backed by cognitive psychology, has suggested links (or patterns) of how people should reason, how people actually reason, and how to generate explanations that support reasoning.

2.3. Design Methods

There is only scarce considerations of design methods for telling designers and developers *how* to design ethical and responsible AI systems using a structured process. At the moment, most of the guidelines mentioned in section 2.1 suggest adopting a User-Centred Design (UCD) process involving user research, designing and prototyping and evaluating, using techniques such as interviews, observations, and user testing. Yet given that many have argued that AI system design pose significant challenges [6, 7, 8], there is yet a dearth of work that addresses design methods that guide designers and developers to develop responsible AI.

Very recently, design methods have been proposed that focus on designing AI algorithms with users. WeBuildAI [22] proposes a framework of steps that involves users in designing algorithms. This method proceeds by investigating feature engineering and selection through surveys and interviews, model building through pair-wise comparison of use by users, and finally model selection through exposing the model decisions.

The most well-known attempt to establish a design method for ethical AI user interfaces is *transparency design* [5]. This work proposes a stage-based process to first investigate mental models of experts and then users to establish a target mental model of what needs to be explained, before iteratively prototyping the user interface to establish how to communicate the explanations and then evaluating it. To develop the mental models of experts, interviews and workshops are suggested, while to investigate users' mental models it is suggested to employ surveys, interviews, task-based studies and drawing tasks. For developing the target model, card sorting, interviews and focus groups were proposed. Designing and evaluating the user interfaces can involve focus groups, workshops, and think-aloud studies. There are now several case studies that have used this process to successfully implement explanations in AI interfaces [23, 24, 25].

Our work is concerned with investigating

[1] http://ui-patterns.com/

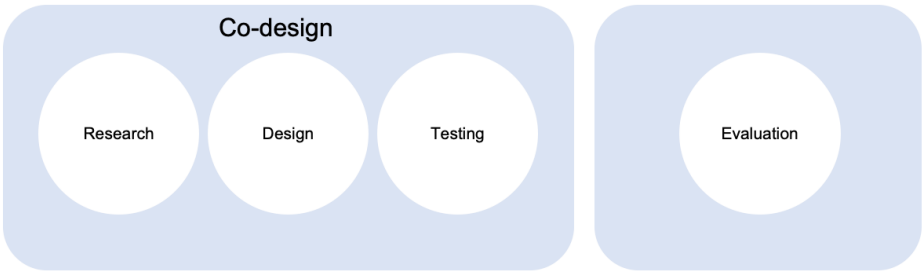

Figure 1: The CoFAIR process

design methods for user interfaces that can help with making the fairness of AI algorithms transparent, and then help with mitigating fairness issues by incorporating user feedback back into the algorithm.

3. The CoFAIR method

We present here our method to Co-design Fair AI InteRactions, CoFAIR (Fig. 1). This method is based on a co-design process [26] which aims to work closely with users to develop solutions through a participatory design approach. As other co-design approaches, It is characterised by very close involvement of a small number of users in all stages of designing a solution, in which these users are empowered to be on equal footing with researcher and designers. Co-design has been successfully adopted to design human-centred technology in other settings [27, 28] however how to use co-design in shaping AI solutions has not been investigated yet.

Our proposed method for responsible AI includes a series of co-design workshops with participants/co-designers that focus on user research, conceptual and detailed design, and initial testing, which is then broadened in a final user evaluation stage.

3.1. Co-design workshops

To start, CoFAIR comprises a series of workshops to work closely with a limited number of participants to research the topic area, to develop some designs, and then to test those designs. To set up these workshops, a number of considerations will need to be made:

3.1.1. Recruitment

Participants in a workshop should be the targeted users of an AI system. The aim is to closely involve these participants in designing a solution that is right *for them*, and to align the design with their requirements. If there are a number of different user groups that are distinct in their background, use cases or tasks, then separate workshops should be organised for them. The users will not need to have a detailed technical understanding or any experience with system design or development, as they will be supported by researchers, designers, and developers. Ideally, they should be relatively representative of the user group in terms of background and demographics. For each workshop, the number of participants should be kept low, between 3-6 people, so as to encourage interaction between participants.

3.1.2. Workshop Aims and Structure

We suggest that the workshops aim to cover three main steps in user-centred design: user research, conceptual and detailed design, and testing. User research in these workshops should investigate the users' current conceptualisations and experience within the topic area, pain points, and high-level needs and wants. This user research can be formalised and communicated through co-created personas that reflect the target user group [27, 29] or could be more informal as simple lists of requirements. Conceptual and detailed design will involve the participants in surfacing what information and interactions are needed to achieve their tasks while also clarifying how to present this in the user interface. This might be documented in storyboards, user journeys, and sketches, or produce scenario-based object-action analyses. Last, these design should be prototyped, either using low-fidelity paper prototypes or more high-fidelity clickable wireframes, and then tested with participants.

Depending on the complexity of what is to be designed, these steps need to be spread over a series of workshops. Most naturally, these steps suggest three sequential workshops, each with a distinct focus on user research, design, and testing. It might be possible to combine user research and design, and thus reduce the number of workshops to two. However, more iterations might be needed to explore design options and iteration of prototypes, and thus more workshops might need to be scheduled. Our method is flexible enough to accommodate this.

3.1.3. Workshop Activities

To achieve the aims of the workshops, co-design usually proceeds with group-based, hands-on activities and discussion around these activities. For user research, these could involve real or hypothetical scenarios and the user experiences around the topic. Activities would typically explore problematic aspects and the challenges that users face in carrying out a user task. They would also probe for basic understandings and conceptualisations around the topic of investigation. These can be (but don't have to be) documented in personas, and recent work has shown how these personas can be co-created with co-design participants [27, 29].

Activities that aim to support design are also kept very concrete. Typically this would investigate a scenario of use, either real or fictitious. As part of conceptual design, participants would usually be invited to go through the scenario of use and indicate what they would look for, what interactions they would expect, and what information the system would need to communicate. It is sometimes helpful to develop storyboards or user journeys with co-design participants. Detailed design can flesh out design options through sketches, however, this needs to be carefully supported and scaffolded as participants are often too timid to sketch themselves.

In testing, a prototype, often created or refined by a designer/developer offline, is exposed to evaluation by co-design participants. Again, a real or fictitious scenario is used to explore how the prototype might be used and what improvements are necessary for a subsequent iteration.

3.2. Broader Evaluation

A common criticism of co-design is the limited number of participants that are involved in developing a solution. This leads to the fear that while the solution is optimally adapted to the 3-6 participants in the workshops, it is unsuitable for the wider user population. Our method suggests that co-design is always followed by broader evaluation of the designed system through evaluations with users.

This can take various forms, such as think-aloud user testing or large-scale crowd-sourced system use.

4. Method Case Study: Loan Application Fairness

In order to show how our method can be instantiated in practice, we present a case study in which we investigated how to develop user interfaces that allow users to explore the fairness of AI loan application decisions. Loan applications decisions are increasingly being automated or supported using AI models (typically, employing logistic regression). This study targeted three different user groups in two iterations: non-expert members of the public (iteration 1), loan officers and data scientists (iteration 2). Iteration 1 details how we instantiated the method with non-expert customers, while iteration 2 is concerned with the method used with loan officers and data scientists. We focus on the techniques employed in our method; we will report on the findings of these studies elsewhere.

4.1. Iteration 1: Non-expert Members of the Public

We ran a series of co-design workshops with a total of 12 participants in the USA, UK, and Japan. Because of COVID-19 restrictions we had to change our planned face-to-face workshops to be conducted entirely online.

4.1.1. Co-Design participants

We recruited 3 participants (2 women, 1 man, mean age 47.3) for the co-design workshops held in the USA, 5 participants (3 women, 2 men, mean age 34.2) in the UK, and 4 participants (3 women, 1 man, mean age 33.75) in Japan through social media and personal contacts. All participants' ethnicities broadly reflected the population of the country, and most participants had been educated to a Bachelor degree level. We paid an incentives of £40, or equivalent in the local currency.

4.1.2. Workshop procedure

For each country, we held 2 co-design workshops; these two workshops were 3 weeks apart. Both workshops lasted 2 hours.

In workshop 1, we conducted user research and conceptual design. For the user research part, we investigated how participants defined fairness, and then how they explored fairness in loan decisions. For investigating how participants viewed AI fairness, we first got participants to tell us about their own experiences of fair or unfair decisions that affected them, especially if they encountered AI in that decision-making. We then also probed them to consider fairness of using AI systems in hiring or making medical decisions and what makes AI systems fair or unfair.

To continue user research and start on conceptual design, we constructed an activity involving four fictitious loan application scenarios (Fig. 2). This allowed us to further investigate what attributes and information they were looking for to assess the fairness of the applications' outcomes and potentially what they would change to make the decisions fairer. Each scenario was discussed in turn, whether it was fair, why (based on the information included in the application or their experience of the decisions they had seen), and what information would have been useful for them to assess fairness better. We changed some of the application scenario details to localize them to each country (e.g. names, currency, dates) but otherwise kept them the same. We showed participants information that is usually collected as part of a loan application process, based on the application form of a well-

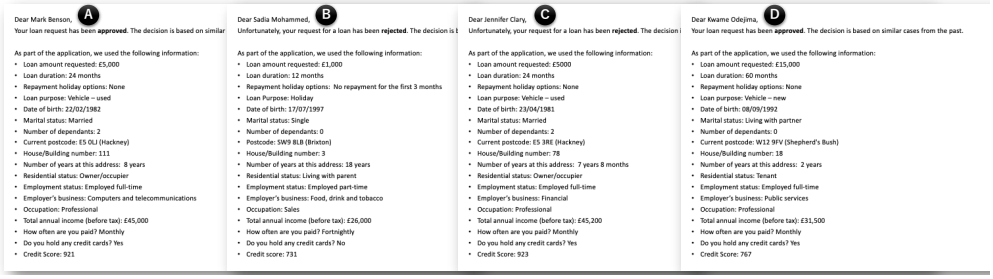

Figure 2: (A) Application 1: Mark Benson, (B) Application 2: Sadia Mohammed, (C) Application 3, Jennifer Clary, and (D) Application 4: Kwame Odejima

known international bank. Application 1 (USA/UK: Mark Benson or Kazufumi Takahashi) was always approved, as it was a 'safe' application, with a homeowner with a very good credit score applying for a small loan to buy a used car. Application 2 (USA/UK: Sadia Mohammed or Chihe Pak) was rejected, as it was a more 'risky' application with low income, part-time job and low credit score. We also included her application to investigate any potential minority or age biases. Application 3 (USA/UK: Jennifer Clary or Maika Suzuki) was also rejected but crucially her details were very similar to Mark Benson. This was to introduce an application that seemed, without any further information, to be blatantly unfair. Finally, application 4 (USA/UK: Kwame Odejima or Dũng Nguyên,) was accepted although it seemed more 'risky'.

After the workshop, two researchers reviewed the workshop recordings and analysed the participants' definitions of AI fairness and how they thought AI could be made fairer. For each scenario, we analysed what criteria they used to assess fairness, how they were using information to explore fairness, and what other information they wanted to be able to assess whether a loan application decision was fair, or potentially biased. Based on this analysis, we constructed clickable wireframes to instantiate their input in an interface. We did this by carefully mapping information that they used for fairness assessments and requests for further information obtained in workshop 1 to interface design elements, and we did not involve participants in detailed design activities.

In workshop 2, we moved on to a testing activity. We structured our discussion on the clickable wireframes, and developed some scenarios to explore fairness using the clickable prototype. Going through each screen's functionality, we discussed what helped to understand if the application decisions were fair, what additional information would they like to determine fairness, and what feedback they would like to give to mitigate fairness.

4.1.3. Broader evaluation

Following the co-design workshops, we implemented an improved interface. We then set up an online study to investigate how this prototype is employed by end-users to assess the fairness of an AI system, and how suggested changes to the model affect fairness.

We recruited 388 participants (129 female, 256 male, 2 Other and 1 preferred not to say) through Prolific[2], an online research platform, and paid them £3.50 for an expected 30-minute session. About half of of our participants had some programming experience and familiarity with AI, machine learning or statistics, and

[2]https://www.prolific.co/

146 participants had at least a Bachelor degree.

We asked participants to interact with the interface to a assess the fairness of an AI system. Instead of using an open-source dataset, the AI system we developed was based on an anonymized loan decisions dataset we obtained from Intesa Sanpaolo. This dataset contains decisions made on 1000 loan applications and has 35 attributes including the label of whether the loan application was accepted or rejected. These attributes include demographic information of the applicant (age, gender, nationality, etc), financial information (household income, insurance, etc), loan information (amount of loan requested, purpose of loan, loan duration, monthly payments, etc), as well as some information of their financial and banking history (years of service with the bank, etc). There were also some attributes that related to internal bank procedures, such as a money laundering check and a credit score developed by the bank. We developed a logistic regression model after removing sparse values, or where multiple attributes had similar values; the accuracy of the resulting model was 0.618. Note that the model was unfair with respect to the nationality attribute: 'foreign' applicants tended to be rejected more frequently than citizens, using disparate impact as a fairness metric.

The evaluation consisted of a brief pre-questionnaire and tutorial, 20 minutes of free use of the interface to assess fairness, and a post-questionnaire. To evaluate the use of this prototype we captured participants' ratings of the AI fairness and key interactions with the user interface where logged. We also asked them to describe in their own words what strategies they used to assess the fairness of the system, any systematic fairness issues they had spotted, and their views on suggesting changes and addressing fairness. We then finished the study by asking them to rate their task load using the NASA-TLX questionnaire [30].

On study completion, we analysed the interactions with the prototype to evaluate whether this prototype was effective in supporting users in exploring the fairness of an AI model.

4.2. Iteration 2: Loan Officer and Data Scientists

4.2.1. Co-design Participants

This iteration was focused on exploring how to support loan officers and data scientists to explore the fairness of loan application decisions. These two stakeholder groups are different: loan officers typically act as intermediaries between the bank and customers and had practical experience of loan decision making, while data scientists have experience in modelling and supporting and/or investigating customer application decisions. For this study, we recruited six loan officers (5 men, 1 woman, mean age 36.5) and six data scientists (3 men, 3 women, mean age 29.7) through Intesa Sanpaolo.

4.2.2. Workshop Procedure

Due to Covid-19 and logistical limitations, all interactions with the users were conducted online. We structured the activities into two workshops, each lasting 2 hours. Both workshops were repeated for each separate stakeholder group.

As with the previous iteration, the aim of workshop 1 was to conduct user research into how fairness was perceived by these user groups, and to carry out initial conceptual design. Workshop 1 started off by discussing the aspects that make decisions in loan applications fair or unfair to get an insight on participants' loan application experience and unfair scenarios that they may have come up against. This was followed by how AI could

impact loan application decision-making and fairness.

To further our user research and also understand what key information is important to use in conceptual design, we then introduced an activity to explore the anonymized loan decisions dataset we obtained from Intesa Sanpaolo. The dataset was sent ahead of the workshop so that participants could have time to look at it and have it available on their computers during the session. The discussion elicited information on participants' process, information needs, and the functionality required to develop an interface. To help participants investigate the dataset, a data visualisation tool was created which was used to present the dataset should participants require it. It provided the ability to slice the features on the fly and present them using various chart types such as histograms, scatter plots, bar graphs and a strip plot.

Next, we introduced an activity to reflect on a causal graph, showing causal relationships between the dataset attributes. This causal graph was derived through automatic discovery, showing how attribute values and the loan application decisions are related to each other. Through this activity we aimed to understand how these users might interpret the causal graph and how this might be employed in exploring the dataset for fairness.

After the first workshops, a researcher analysed the audio recordings to derive findings about how these user groups judged whether loan applications were fair, how these users explored the dataset to determine fairness, and how they interpreted the causal graph. Based on this analysis the researcher developed a clickable wireframe to be used in workshop 2 (Fig. 3). Again, we did not involve the users in detailed design. Due to implementation constraints, we only made a selection of the wireframe interactive, and focused on a scenario in which to explore the relationships between citizenship, gender, credit risk level, loan amount and number of instalments in detail.

The aim of workshop 2 was to informally test the clickable wireframe. This wireframe was screen shared and the researcher 'drove' the interactions with it and acted as an extension on the participants' behalf, clicking through it. The researcher then stepped through it with the respective user groups, and probed whether they understood how it worked, wheth-er the information was useful for exploring fairness, or what could be improved.

Analysis of the second workshop investigated changes that needed to be made to improve the clickable prototype for broader evaluation. Based on this analysis the researchers designed a prototype (Fig. 4).

4.2.3. Broader Evaluation

The evaluations were conducted as one-to-one user tests, unlike the workshops in the previous phase. A total of 17 participants were recruited through Intesa Sanpaolo: 8 loan officers (5 men, 3 women, mean age 38) and 9 data scientists (5 men, 4 women, mean age 31.8). All participants held a master's degree or higher.

We developed ten tasks for participants to go through the prototype, from setting up the dataset to explore to investigating the dataset, using different components of the user interface. The study concluded with a post-questionnaire used to evaluate users' experience. This questionnaire comprised ratings aimed at quantifying how effective the prototype was in supporting users in assessing fairness including information, functionality and reasoning, free comments to express their feedback about the prorotype, and the NASA TLX questionnaire [30].

The broader evaluation was analysed as to what worked well and what did not, in order to develop functioning interfaces in future.

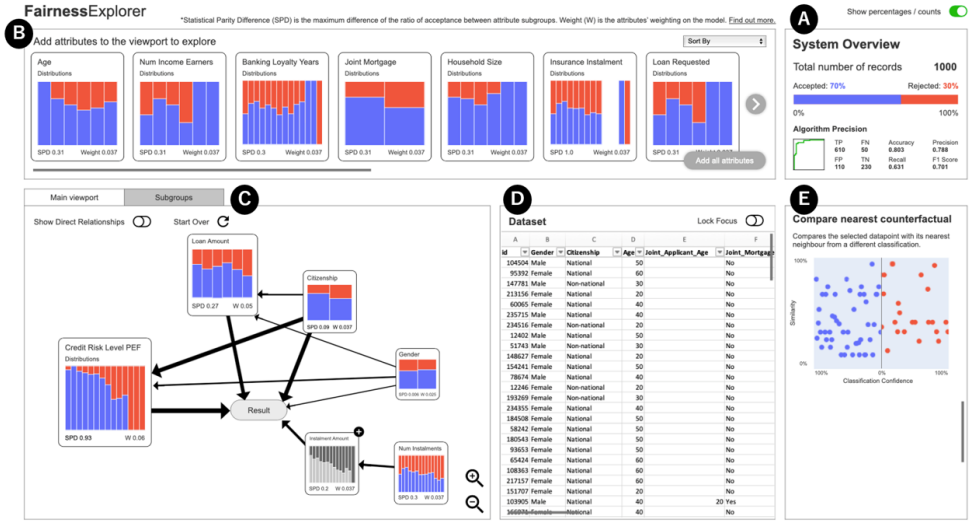

Figure 3: The clickable wireframe used in Workshop 2. (A) System overview. (B) Attribute information including name, value distribution, fairness metric and weight. (C) Causal graph of selected attributes. (D) Dataset. (E) Comparison of currently selected application in dataset and all other applications with respect to similarity and application outcome.

5. Discussion

We have gained some experience from applying co-design in other application domains, and through a case study where we implemented the CoFAIR method to develop interfaces for exploring fairness. This showed that this method can be successfully employed to design interfaces for responsible AI systems. However, we encourage other researchers and practitioners to adopt this method and generate more data points to improve this approach, and also to validate it. In addition, CoFAIR was so far employed under COVID-19 restrictions which meant that all workshop activities and testing had to be conducted remotely online, which impacted what we were able to do. If we had not been placed in this situation, we would have made different choices as how to conduct the workshops. First, due to the online nature we shortened the co-design activities and compressed them into two workshops of two hours each. Ideally we would like to extend them to span three workshops and for a longer duration. Second, facilitation of online discussions is very difficult, and ideally we would have brought users together to discuss this more freely face-to-face. Last, we would have liked to involve users much more in conceptual and detailed design, for example, through sketching or paper prototyping but this is very difficult to do virtually.

We can also note some general limitations of the CoFAIR method which should be considered before it is chosen as a design approach. First, as with all co-design there is a danger that interfaces are developed that only fit the small number of people that were involved as users in the workshops. This can be alleviated through conducting broader evaluations that ensure that the designs are fit for pur-

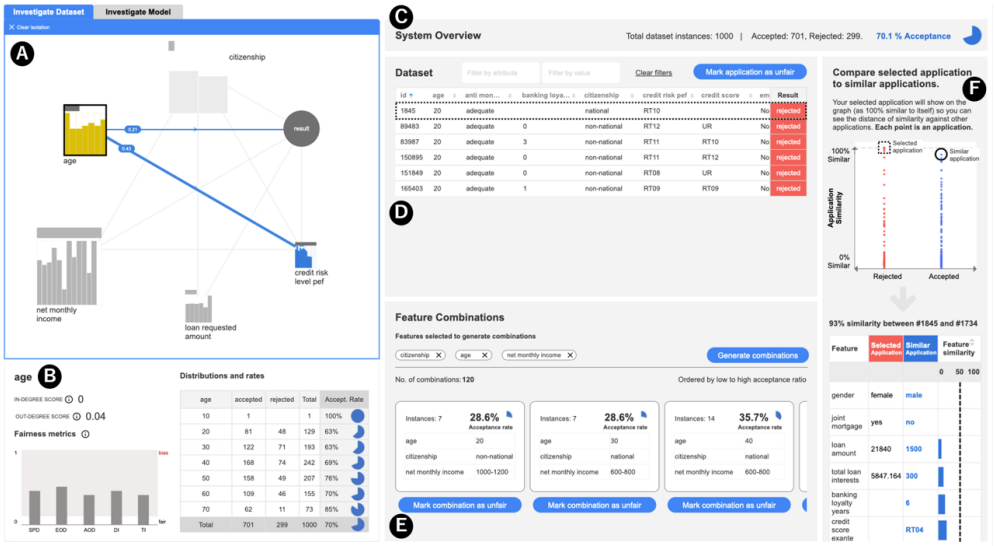

Figure 4: The evaluation prototype. (A) Causal graph of selected attributes. (B) Attribute information including fairness metrics, and value distributions with respect to application decision. (C) System overview including total acceptance. (D) Datset view, highlighting currently selected attribute. (E) Attribute combination view, showing subset information. (F) Comparison of similarity currently selected application with other applications.

pose. Second, it is not a 'discount' methodology that is fast and easy to apply. Implementing it requires several lengthy workshops with users to be organised, separated in time so that researchers and designers can analyse and produce new materials in subsequent activities. This means that even relatively small projects can spread over several months, from initial recruitment of users to a fully refined and evaluated interface. Because we want to guard against 'overfitting' designs to small numbers of participants, it is not advisable to cut short this process and skip the broader evaluation to save on time. Last, this method focuses very much on the mental model of users and does not account for the input of 'experts' or consider how people should reason. Hence, it is possible that we might build in possible biases that users have back into these interfaces, and only support current ways of working. How to successfully mitigate fairness issues, especially through a human-in-the-loop approach, is still an open research question.

We believe that our method is another step to strengthen the design of responsible and ethical AI. A major advantage of CoFAIR is that it produces designs and interfaces that focus heavily on what specific target users need and want. It thus produces 'shrink-wrapped' interfaces that should be eminently suitable to communicate with a specific user group. Taken together, this method could be easily extended to investigate what and how to explain machine learning systems, in order to design more responsible and ethical AI systems.

6. Conclusion

In this paper, we outlined that practical design methods that translate general guidelines into concrete processes to follow are in short supply. We presented the CoFAIR method to design responsible AI: co-design workshops that focus on user research, conceptual and detailed design and initial testing are followed by broader evaluation. We showed how we implemented this method through a case study which focused on supporting non-expert 'end-users', loan officers, and data scientists to explore fairness in loan application decisions. We discussed the considerations that need to be made when choosing this method. We believe that other researchers, designers and practitioners of responsible AI systems can adopt this approach to develop suitable interfaces.

References

[1] M. K. Lee, Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management, Big Data & Society 5 (2018) 2053951718756684. URL: https://doi.org/10.1177/2053951718756684. doi:10.1177/2053951718756684, publisher: SAGE Publications Ltd.

[2] M. Veale, M. Van Kleek, R. Binns, Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–14. URL: https://doi.org/10.1145/3173574.3174014. doi:10.1145/3173574.3174014.

[3] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine Intelligence 1 (2019) 389–399. URL: https://www.nature.com/articles/s42256-019-0088-2. doi:10.1038/s42256-019-0088-2, number: 9 Publisher: Nature Publishing Group.

[4] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing Theory-Driven User-Centric Explainable AI, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, ACM, New York, NY, USA, 2019, pp. 601:1–601:15. URL: http://doi.acm.org/10.1145/3290605.3300831. doi:10.1145/3290605.3300831, event-place: Glasgow, Scotland Uk.

[5] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, H. Hussmann, Bringing Transparency Design into Practice, in: 23rd International Conference on Intelligent User Interfaces, IUI '18, ACM, New York, NY, USA, 2018, pp. 211–223. URL: http://doi.acm.org/10.1145/3172944.3172961. doi:10.1145/3172944.3172961.

[6] G. Dove, K. Halskov, J. Forlizzi, J. Zimmerman, Ux design innovation: Challenges for working with machine learning as a design material, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 278–288. URL: https://doi.org/10.1145/3025453.3025739. doi:10.1145/3025453.3025739.

[7] Q. Yang, A. Steinfeld, C. Rosé, J. Zimmerman, Re-examining whether, why, and how human-ai interaction is uniquely difficult to design, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–13. URL: https://doi.org/

10.1145/3313831.3376301. doi:10.1145/3313831.3376301.

[8] Q. Yang, A. Steinfeld, C. Rosé, J. Zimmerman, Re-examining whether, why, and how human-ai interaction is uniquely difficult to design, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–13. URL: https://doi.org/10.1145/3313831.3376301. doi:10.1145/3313831.3376301.

[9] L. E. Holmquist, Intelligence on Tap: Artificial Intelligence As a New Design Material, interactions 24 (2017) 28–33. URL: http://doi.acm.org/10.1145/3085571. doi:10.1145/3085571.

[10] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for Human-AI Interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–13. URL: https://doi.org/10.1145/3290605.3300233. doi:10.1145/3290605.3300233.

[11] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 220–229. URL: https://doi.org/10.1145/3287560.3287596. doi:10.1145/3287560.3287596.

[12] People + AI Guidebook, 2019. URL: https://pair.withgoogle.com/guidebook.

[13] F. Rossi, A. Sekaran, J. Spohrer, R. Caruthers, Everyday Ethics for Artifical Intelligence, 2019. URL: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

[14] Q. V. Liao, D. Gruen, S. Miller, Questioning the AI: Informing Design Practices for Explainable AI User Experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–15. URL: https://doi.org/10.1145/3313831.3376590. doi:10.1145/3313831.3376590.

[15] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, 2019.

[16] M. A. Madaio, L. Stark, J. Wortman Vaughan, H. Wallach, Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–14. URL: https://doi.org/10.1145/3313831.3376445. doi:10.1145/3313831.3376445.

[17] T. Kulesza, M. Burnett, W.-K. Wong, S. Stumpf, Principles of Explanatory Debugging to Personalize Interactive Machine Learning, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15, ACM, New York, NY, USA, 2015, pp. 126–137. URL: http://doi.acm.org/10.1145/2678025.2701399. doi:10.1145/2678025.2701399.

[18] V. Bellotti, K. Edwards, Intelligibility and Accountability: Human Considerations in Context-aware Systems, Hum.-Comput. Interact. 16 (2001) 193–

212. URL: http://dx.doi.org/10.1207/S15327051HCI16234_05. doi:10.1207/S15327051HCI16234_05.

[19] B. Y. Lim, A. K. Dey, Investigating Intelligibility for Uncertain Context-aware Applications, in: Proceedings of the 13th International Conference on Ubiquitous Computing, UbiComp '11, ACM, New York, NY, USA, 2011, pp. 415–424. URL: http://doi.acm.org/10.1145/2030112.2030168. doi:10.1145/2030112.2030168.

[20] B. Y. Lim, A. K. Dey, Toolkit to Support Intelligibility in Context-aware Applications, in: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, UbiComp '10, ACM, New York, NY, USA, 2010, pp. 13–22. URL: http://doi.acm.org/10.1145/1864349.1864353. doi:10.1145/1864349.1864353.

[21] B. Y. Lim, A. K. Dey, D. Avrahami, *Why and why not* explanations improve the intelligibility of context-aware intelligent systems, ACM, Boston, MA, USA, 2009, pp. 2119–2128. URL: http://portal.acm.org/citation.cfm?id=1518701.1519023Źcoll=portalŹdl=ACMŹtype=seriesŹidx=SERIES260Źpart=seriesŹWantType=ProceedingsŹtitle=CHIŹCFID=31206243ŹCFTOKEN=35340577. doi:10.1145/1518701.1519023.

[22] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, A. D. Procaccia, Webuildai: Participatory framework for algorithmic governance, Proc. ACM Hum.-Comput. Interact. 3 (2019). URL: https://doi.org/10.1145/3359283. doi:10.1145/3359283.

[23] C. Tsai, P. Brusilovsky, Designing Explanation Interfaces for Transparency and Beyond, in: Algorithmic Transparency in Emerging Technologies, Los Angeles, 2019. URL: http://ceur-ws.org/Vol-2327/.

[24] M. Ribera, A. Lapedriza, Can we do better explanations? A proposal of User-Centered Explainable AI, in: Explainable Smart Systems (ExSS), 2019, p. 7. URL: http://ceur-ws.org/Vol-2327/.

[25] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, H. Hussmann, I Drive - You Trust: Explaining Driving Behavior Of Autonomous Cars, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, ACM, New York, NY, USA, 2019, pp. LBW0163:1–LBW0163:6. URL: http://doi.acm.org/10.1145/3290607.3312817. doi:10.1145/3290607.3312817, event-place: Glasgow, Scotland Uk.

[26] E. B.-N. Sanders, P. J. Stappers, Co-creation and the new landscapes of design, CoDesign 4 (2008) 5–18. URL: http://dx.doi.org/10.1080/15710880701875068. doi:10.1080/15710880701875068.

[27] A. Bourazeri, S. Stumpf, Co-designing Smart Home Technology with People with Dementia or Parkinson's Disease, in: Proceedings of the 10th Nordic Conference on Human-Computer Interaction, NordiCHI '18, ACM, New York, NY, USA, 2018, pp. 609–621. URL: http://doi.acm.org/10.1145/3240167.3240197. doi:10.1145/3240167.3240197, event-place: Oslo, Norway.

[28] S. Wilson, A. Roper, J. Marshall, J. Galliers, N. Devane, T. Booth, C. Woolf, Codesign for people with aphasia through tangible design languages, CoDesign 11 (2015) 21–34. URL: http://dx.doi.org/10.1080/15710882.2014.997744. doi:10.1080/15710882.2014.997744.

[29] T. Neate, A. Bourazeri, A. Roper, S. Stumpf, S. Wilson, Co-Created

Personas: Engaging and Empowering Users with Diverse Needs Within the Design Process, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, ACM, New York, NY, USA, 2019, pp. 650:1–650:12. URL: http://doi.acm.org/10.1145/3290605.3300880. doi:10.1145/3290605.3300880, event-place: Glasgow, Scotland Uk.

[30] S. G. Hart, L. E. Staveland, Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, in: Peter A. Hancock and Najmedin Meshkati (Ed.), Advances in Psychology, volume Volume 52 of *Human Mental Workload*, North-Holland, 1988, pp. 139–183. URL: http://www.sciencedirect.com/science/article/pii/S0166411508623869.